논문 2005-42TE-1-3

k-clustering 부공간 기법과 판별 공통벡터를 이용한 고립단어 인식

(Isolated Word Recognition Using k-clustering Subspace Method and Discriminant Common Vector)

남 명 우*

(Myung-Woo Nam)

요 약

본 논문에서는 M. Bilginer 등이 제안한 CVEM(common vector extraction method)을 이용하여 한국어 화자독립 고립단어 인식실험을 수행하였다. CVEM은 학습용 음성신호들로부터 공통된 특징의 추출이 비교적 간단하고, 많은 계산 량을 필요로 하지 않을 뿐만 아니라 높은 인식 결과를 보여주는 알고리즘이다. 그러나 학습 음성의 개수를 일정 한도 이상으로 늘릴 수 없고, 추출된 공통벡터들 간의 구별정보(discriminant information)를 가지고 있지 않다는 문제점을 가지고 있다. 임의의 음성군으로부터 최적의 공통벡터를 추출하기 위해서는 다양한 음성들을 학습에 사용해야만 하는데 CVEM은 학습용 음성 개수에 제한이 있으므로 지속적인 인식률 향상을 기대하기 어렵다. 또한 공통벡터들 간의 구별정보 부재는 단어 결정에 있어서 치명적인 오류의 원인이 될 수 있다. 본 논문에서는 CVEM이 가지고 있는 이러한 문제점들을 보완하면서 인식률을 향상시킬 수 있는 새로운 방법인 KSCM(k-clustering subspace method)과 DCVEM(discriminant common vector extraction method)을 제안하였고 이 방법을 사용하여 고립단어를 인식하였다. 그리고 제안한 방법들의 우수성을 입증하기 위해 ETRI에서 제작한 음성 데이터베이스를 사용, 다양한 방법으로 실험을 수행하였다. 실험 결과 기존 방법의 문제점들을 모두 극복할 수 있었을 뿐 아니라 기존에 비해 계산량의 큰 증가 없이 향상된 결과를 얻을 수 있었다.

Abstract

In this paper, I recognized Korean isolated words using CVEM which is suggested by M. Bilginer et al. CVEM is an algorithm which is easy to extract the common properties from training voice signals and also doesn't need complex calculation. In addition, CVEM shows high accuracy in recognition results. But, CVEM has couple of problems which are impossible to use for many training voices and no discriminant information among extracted common vectors. To get the optimal common vectors from certain voice classes, various voices should be used for training. But CVEM is impossible to get continuous high accuracy in recognition because CVEM has a limitation to use many training voices and the absence of discriminant information among common vectors can be the source of critical errors. To solve above problems and improve recognition rate, k-clustering subspace method and DCVEM suggested. And did various experiments using voice signal database made by ETRI to prove the validity of suggested methods. The result of experiments shows improvements in performance. And with proposed methods, all the CVEM problems can be solved with out calculation problem.

Keywords: 고립단어 인식(isolated word recognition), 공통벡터(common vector), 부공간(subspace), 2차 이산 코사인 변환(2 dimension discrete cosine transform), 청각모델(auditory model)

I. 서 론

음성신호는 화자의 성별, 나이, 출생지, 주위 잡음, 정

* 정회원, 혜전대학 디지털전자디자인과
(Dept. of Digital Electro-Design, Hyejeon College)
접수일자: 2004년11월19일, 수정완료일: 2004년12월29일

신적 상태, 발성기관의 구조 등 다양한 정보를 포함하고 있다. 그리고 이런 정보들은 개인의 독특한 특징들을 나타내는 것이므로 화자인식과 같은 분야에 응용될 수 있지만 또 다른 의미에서는 공통된 특징 추출을 어렵게 한다. 음성신호는 주파수영역 또는 시간영역에서 다양한 파라미터들로 표현될 수 있으며 이러한 파라미터들은 모두 벡터로 변환될 수 있다. 이렇게 벡터로 표현된 음성신호는 음성인식과 화자인식 분야에서 인식률을 높이기 위한 방법으로 많이 사용되고 있다. 벡터로 표현된 음성신호는 일반적으로 KLT(Karhunen-Loève Transformation)방법을 사용하여 직교(orthogonal) 파라미터로 변환한 후, DTW(Dynamic Time Warping), NN(Neural Networks), HMM(Hidden Marcov Model)과 같은 인식기에 입력으로 사용된다. 벡터를 이용한 또 다른 음성신호처리 기법으로 M. Bilginer 등이 제안한 공통벡터 추출방법(Common Vector Extraction Method)이 있다[1]. CVEM은 같은 단어군에 속한 음성신호의 차들(화자들의 특징과 잡음신호의 합으로 이루어짐)로 정규직교 벡터공간(orthonormal vector space) $\mathbf{B}$를 만든 후, 입력으로 들어온 음성신호를 벡터공간 $\mathbf{B}$에 사상시켜 공통된 성분을 얻어내는 방법인데, 공통된 성분을 추출하는 과정동안 고유값과 고유벡터의 계산을 필요로 하지 않으며 모든 불필요한 정보들(주위 잡음, 화자 특성 등)을 효과적으로 제거할 수 있는 장점을 가지고 있다. 이는 입력 음성신호가 화자들의 특징과 잡음신호들의 합으로 이루어진 벡터공간 $\mathbf{B}$로 사상될 때 불필요한 정보들이 제거되는 효과를 얻을 수 있기 때문이다. 따라서 CVEM은 학습(training)용 음성신호들로부터는 최적의 공통벡터를 추출해낼 수 있으며 100% 인식률을 얻을 수 있다. 그러나 CVEM은 음성신호의 차들로부터 정규직교 벡터공간 $\mathbf{B}$를 구성할 때 학습용 음성신호들이 서로 일차독립이어야 한다는 제한조건이 따른다[1]. 이 제한조건은 임의의 음성군으로부터 최적의 공통벡터를 추출할 때 다양한 음성들을 학습에 사용하는 것을 불가능하게 하며 학습용 음성신호들의 수가 증가하더라도 지속적인 인식률 향상을 할 수 없게 한다. 즉, 음성신호들이 n-차원의 벡터들로 구성되어 있다면 각 단어군은 n개를 초과하는 학습용 음성신호를 사용할 수 없다는 것이다. 이와 같은 경우는 n-차원의 벡터들로 구성된 임의의 단어군에 n개를 초과하는 학습용 음성 m개가 존재할 경우 $(m-n)$개의 학습용 음성은 공통벡터 추출에 전혀 영향을 주지 못하는 결과를 초래한다. 또한 CVEM은 추출된 공통벡터들 간의 구별정보(discriminant information)를 명확하게 정의하고 있지 못하는 문제점을 가지고 있다.

공통벡터들 간의 구별정보 부재는 각 단어군이 서로 작은 구별정보(discriminant information)를 가지고 있을 경우 단어 결정부에서 치명적인 오류의 원인이 될 수 있다[2]. 본 논문에서는 CVEM이 가지고 있는 단점을 보완하면서 성능을 향상시킬 수 있는 새로운 두 가지 방법을 제안하였다. 먼저 첫 번째로 제안한 방법은 k-clustering subspace method(KCSM)이다. KCSM은 CVEM이 가지는 벡터의 차원이라는 제한조건을 해결할 수 있는 방법으로, 학습용 음성의 증가와 관계없이 공통벡터 추출을 가능하게 한다. 또한 전체 계산량에도 큰 차이가 없는 장점을 가지고 있다. 다음으로 제안한 방법은 각 단어군 간의 분별력 정보를 증가시킬 수 있는 discriminant 공통벡터 추출방법이다.

이 방법은 인식의 후반부에서, 얻어진 discriminant 공통벡터간의 Euclidean 거리와 CVEM으로 얻어진 공통벡터간의 Euclidean 거리를 각각 결합하여 단어를 인식하는데 사용하는 것으로, 실험 결과 더 높은 인식률을 얻을 수 있었다.

Ⅱ. Bilginer의 공통벡터

1. 공통벡터 추출방법

공통벡터 추출을 위해 몇 가지 수식을 정의한다.

$\mathbf{R}^n$: n-차원의 벡터공간

$< \boldsymbol{x}, \boldsymbol{y} >$: 벡터 $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^n$ 의 내적

$| x | = < x, x >^{1/2}$: 벡터 $\boldsymbol{x}$의 Euclidean norm

n개의 성분을 가지는 일차독립인 벡터들 $\boldsymbol{a_1}, \boldsymbol{a_2}, \cdots, \boldsymbol{a_m} \in \mathbf{R}^n$을 가정해 보자. 단, $m < n$을 만족하며 벡터 $\boldsymbol{a}_i (i = 1, 2, \cdots, m)$는 각각 다른 화자로부터 발성된 학습 음성들로 모두 같은 단어군(class)에 속한다고 가정하자. 이러한 가정을 통해서 벡터 $\boldsymbol{a_i}$는 아래와 같이 표시될 수 있다.

$$\begin{aligned} a_1 &= X + a_{1,diff} \\ a_2 &= X + a_{2,diff} \\ &\vdots \\ a_m &= X + a_{m,diff} \end{aligned} \tag{1}$$

위 식에서 X는 a_i가 속한 단어군이 가지는 공통된 성분을 뜻하며, $a_{i,diff}$는 i번째 화자의 특징과 잡음 신호의 합을 의미한다. 식 2는 참조벡터와 나머지 단어들과의 차로 만들어진 벡터 $b_i\,(i=1,2,\cdots,m-1)$를 구하는 과정이다.

$$\begin{aligned} b_1 &= a_2 - a_1 \\ b_2 &= a_3 - a_1 \\ &\vdots \\ b_{m-1} &= a_m - a_1 \end{aligned} \tag{2}$$

b_i는 서로 일차독립인 특성을 가지며, 임의의 단어군에서 공통성분을 제외하고 남은 화자들의 특징과 잡음 신호의 합들로 구성된 벡터들이다. b_i에 Gram-Schmidt의 직교화 방법(orthogonalization method)을 사용하여 만든 정규직교벡터(orthonormal vector)를 z라고 정의하자. z는 b_i의 직교 기저(orthogonal basis)를 이루며, $z = \{z_1, z_2 \cdots, z_{m-1}\}$ 와 $< z_i, z_j > = \delta_{ij} = \{1 \text{ if } i=j;\ 0 \text{ if } i \neq j\}$ 의 특성을 가진다. 식 1과 식 2, 그리고 z를 이용하여 식 4를 유도할 수 있다.

$$\begin{aligned} \tilde{a}_i &= \widetilde{a_{common}} \\ &= a_i - \bar{a}_i = a_1 - \bar{a}_1 \quad (i=1,2,\cdots,m) \end{aligned} \tag{3}$$

$$\begin{aligned} \bar{a}_i = &< a_i, z_1 > z_1 + < a_i, z_2 > z_2 \\ &+ \cdots + < a_i, z_{m-1} > z_{m-1} \end{aligned} \tag{4}$$

식 3에서 얻어진 $\tilde{a}_i$는 i에 관계없이 일정하고, 임의의 단어군에 속한 벡터 a_i들의 공통된 성분을 의미한다. 즉, $\tilde{a}_i$는 벡터 a_i와 부공간 $\mathbf{B}$로 사상된 a_i의 사상벡터 $\tilde{a}_i$의 차로 이루어진 부공간 $\mathbf{B}$와 수직을 이루는 벡터를 의미한다.

$$\begin{aligned} F_2(\widetilde{a_{common}}) &= \{\| a_1 - X - \bar{a}_1 \|^2 + \cdots \} \\ &\quad + \| a_m - X - \bar{a}_m \|^2_{X=\widetilde{a_{common}}} \\ &= \| a_1 - \widetilde{a_{common}} - \bar{a}_1 \|^2 + \cdots \\ &\quad + \| a_m - \widetilde{a_{common}} - \bar{a}_m \|^2 \end{aligned} \tag{5}$$

위의 식은 식 3을 이용하면 $F_2(\tilde{a}_{common}) = 0 + \cdots + 0 = 0$을 얻을 수 있다. 따라서 F_2는 $X_{opt} = \tilde{a}_{common}$일 때 가장 작은 F_2를 얻을 수 있으며, 이때의 F_2값은 학습 음성들에 대해 영이 된다. 식 3의 공통된 성분, 즉 공통벡터는 각각의 단어군에 대해 하나씩 생성되며 임의의 입력 음성을 판단하기 위한 참조벡터로 사용된다.

공통벡터를 이용하여 임의의 입력 음성을 인식하는 방법은 다음과 같다. 인식 실험을 위한 임의의 입력 음성들 중 같은 단어군에 속한 단어들을 c로 가정하자. 단, $c = \{c_1, c_2, \cdots, c_s\} \in \mathbf{R}^n$이며, c_i는 서로 일차독립이고 $s < n$을 만족한다. 이때 최적화 함수 F_2는 아래와 같다.

$$\begin{aligned} F_2 &= \| c_{1,diff} - \bar{c}_1 \|^2 + \cdots\cdots + \| c_{s,diff} - \bar{c}_s \|^2 \\ &= \| c_1 - \bar{c}_1 - X \|^2 + \cdots + \| c_s - \bar{c}_s - X \|^2 \\ &= \| \tilde{c}_{remaining} - X \|^2 + \cdots + \| \tilde{c}_{remaining} - X \|^2 \end{aligned} \tag{6}$$

$$\tilde{c}_{remaining} = c_i - \bar{c}_i = c_i - \sum_{j=1}^{m-1} \langle c_i, z_j \rangle z_j \tag{7}$$

추출되는 공통벡터는 F_2가 항상 최소가 되도록 결정되기 때문에 이러한 특성을 이용한 단어 결정식은 식 8과 같다. w는 학습에 사용된 단어군의 총 개수이며 $\tilde{a}^k_{common}$는 각 단어군으로부터 추출된 공통벡터이다. 그리고 i는 결정식으로부터 얻어진 단어군을 의미한다.

$$i = index\left[\min_{1 \le k \le w}\left(\left\| \tilde{c}_{remaining} - \tilde{a}^k_{common} \right\|\right)\right] \tag{8}$$

위에서 유도된 식들을 좀 더 일반 확장시켜 보면 식 9와 10은 항상 성립되며, m은 공통벡터 추출을 위한 임의의 단어군에 속한 단어의 개수이다.

$$\boldsymbol{F}_2(x)\,|_{x=\tilde{a}_{common}} \leq \boldsymbol{F}_1(x)\,|_{x=a_{ar}} \tag{9}$$

$$\boldsymbol{F}_2^{m+1} \leq \boldsymbol{F}_2^{m} \tag{10}$$

2. 공통벡터 추출의 문제점

CVEM은 공통성분 추출의 간편함과 적은 계산량만으로 높은 인식률을 얻을 수 있다는 장점에도 불구하고 몇 가지 중요한 문제점들을 내포하고 있다. 이러한 문제점들은 CVEM이 다양한 제품으로 실용화되는데 큰 걸림돌로 작용할 수 있다. CVEM이 가지고 있는 문제점은 두 가지로 나누어 생각할 수 있으며 인식 성능에 많은 영향을 미친다.

가. 문제점 1

CVEM은 학습용 벡터의 차원이 임의의 단어군에 속한 학습용 음성의 개수보다 커야 한다는 제한조건을 필요로 한다. 이는 학습 음성의 수가 벡터의 차원을 넘을 경우 각 벡터들이 일차독립이라는 가정을 만족할 수 없기 때문이다[3]. 이러한 제한조건은 벡터의 차원을 증가시키는 방법으로 해결할 수도 있으나 다양한 음성신호로부터 공통된 특성을 추출하기 위해 벡터의 차원을 무한히 늘리는 것은 불가능한 일이다. 왜냐하면 벡터의 차원이 증가함에 따라 공통성분을 추출하기 위한 연산량이 비례하여 증가할 뿐만 아니라 인식기의 구현을 위한 시스템 사양이 높게 요구되기 때문이다. 그러나 음성은 화자의 성별, 나이 감정 등에 따라 다양한 특성을 가지므로 가능한 많은 음성을 학습용으로 사용하는 것이 바람직하다. 이와 같은 이유로 CVEM은 일정한도 이상의 인식률 향상을 기대하기 어렵다.

나. 문제점 2

각 단어군에서 추출된 공통벡터들간의 구별정보가 전혀 정의되어 있지 않다. 즉, 각 단어군에서 추출된 공통벡터들은 학습에 사용된 음성신호들을 이용하여 실험할 경우 100%의 인식률을 얻을 수 있지만, 임의의 입력음성에 대해서는 각 단어군에서 추출한 전체 공통벡터와의 우도(likelihood)를 계산하여 인식단어를 결정하게 된다. 그런데 만약 각 단어군의 공통벡터간 구별정보(discriminant information)가 작을 경우 이는 오인식(recognition error)의 주된 원인이 될 수 있다[2][4]. 따라서 이러한 문제를 해결하기 위해 만약 각 단어군의 공통벡터들간 구별정보를 크게 할 수 있다면, 오인식을 줄일 수 있을 뿐 아니라 이는 곧 인식률의 향상으로 이어질 것이다.

Ⅲ. KCSM와 DCVEM을 결합한 공통벡터

1. k-clustering subspace method(KCSM)

CVEM이 가지는 공통성분 추출의 간편함과 적은 계산량의 장점을 살리면서 CVEM이 가지고 있는 문제점을 해결하기 위하여 본 논문에서는 k-clustering subspace method (KCSM)를 제안하였다. KCSM은 학습 음성의 수에 무관하게 공통벡터 추출을 가능하게 하는 방법으로, 기존 방식보다 연산량의 증가가 크지 않으며, 다양한 음성신호들을 학습할 수 있게 해주는 장

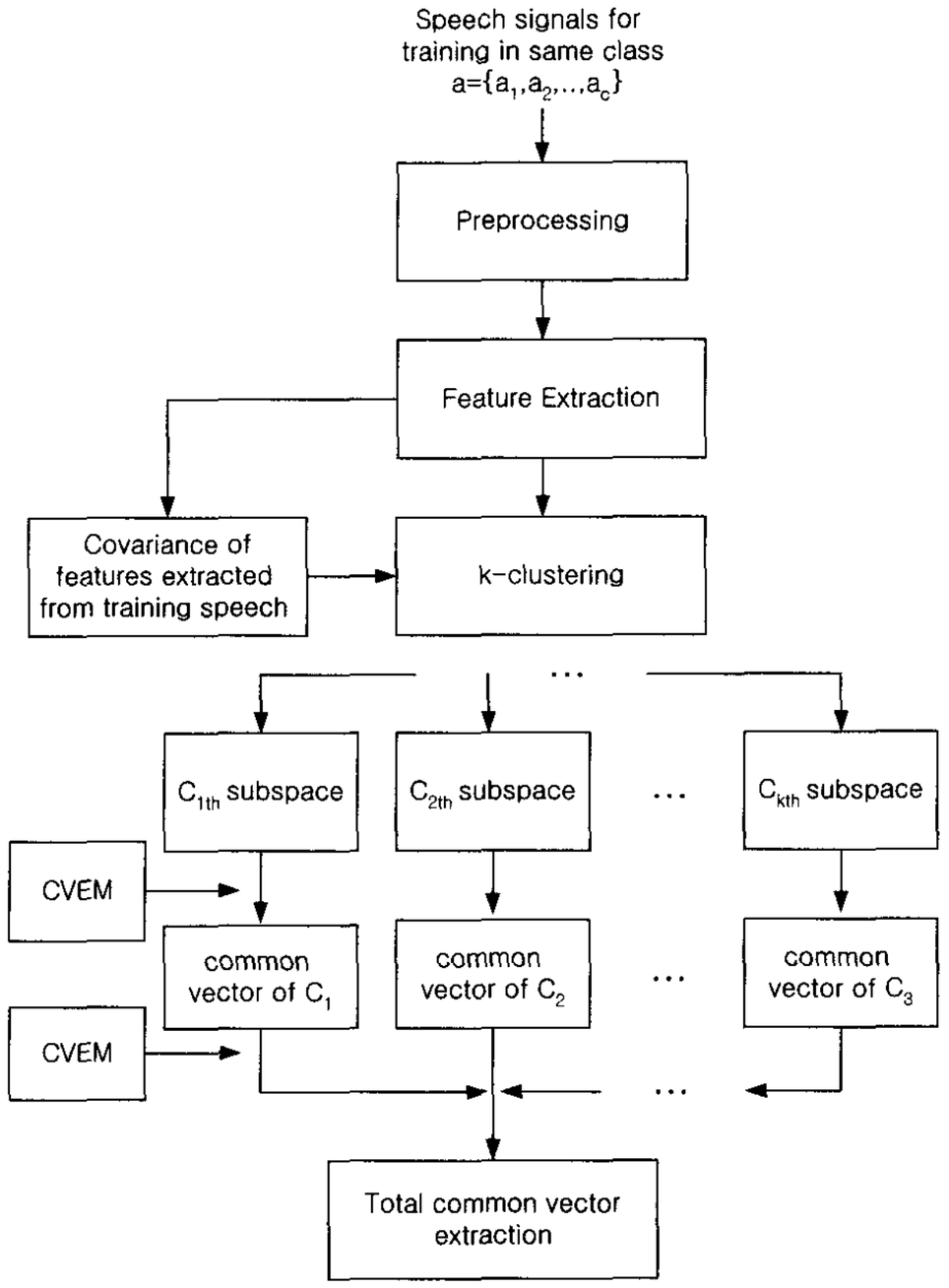

그림 1. KCSM의 구성도
Fig. 1. Block diagram of KCSM.

점을 가진다. 연산량의 경우 CVEM은 모든 직교단위벡터와의 내적을 계산해야 하지만 KCSM은 각 소그룹을 구성하고 있는 직교단위벡터들과만 내적을 계산하면 된다. 그리고 각 소그룹을 구성하고 있는 직교단위벡터 개수의 합은 전체 직교단위벡터의 개수와 동일하다. 따라서 KCSM는 우도를 계산하는 부분을 제외하고는 CVEM과 동일한 연산량을 사용하게 된다.

벡터$\boldsymbol{a_1}, \boldsymbol{a_2}, \cdots, \boldsymbol{a_c} \in \mathbf{R}^n$ (단, $c \gg n$)을 가정해 보자. 벡터 $\boldsymbol{a}_i$ (단, $i=1,2,\cdots,c$)는 각각 다른 화자가 발성한 학습 음성들로서 모두 같은 단어군에 속한다고 가정하자. 먼저 벡터 $\boldsymbol{a}_i$들을 k (단, $k < c$)개의 소그룹으로 묶는다. 단, 각 소그룹에 속하는 학습 음성들의 개수는 n을 넘지 않도록 제한조건을 두어야 한다. 소그룹으로 묶는 방법은 여러 가지가 있을 수 있으나 실험 결과 본 논문에서 제안한 공분산을 이용한 방법이 안정된 인식률을 보여 주었다. 다음으로 k개의 소그룹들로부터 기존의 공통벡터 추출방법(CVEM)을 사용하여 공통성분을 추출한다. k개의 소그룹에서 모두 k개의 공통벡터가 추출될 것이며 각 공통벡터의 추출에 사용된 $\boldsymbol{k}$개의 정규직교벡터 집합들도 생성될 것이다. 마지막 과정으로 $\boldsymbol{k}$개의 공통벡터들을 이용하여 다시 전체 공통벡터를 추출한다. 추출된 전체 공통벡터는 벡터 $\boldsymbol{a}_i$에 포함된 최적의 공통된 성분을 의미한다.

2. 판별(discriminant) 공통벡터 추출방법 (DCVEM)

임의의 입력 음성에 대해서는 각 단어군에서 추출한 전체 공통벡터와의 우도를 계산하여 인식단어를 결정하게 된다. 이때 각 단어군의 전체 공통벡터간 구별정보(discriminant information)가 작을 경우 오인식의 원인이 될 수 있다. 따라서 만약 각 단어군의 전체 공통벡터들간 구별정보를 크게 한다면, 오인식을 줄일 수 있을 뿐 아니라 인식률의 향상으로 이어질 것이다. 먼저 $\boldsymbol{m}$개의 벡터들로 구성된 T개의 단어군을 가정해 보자. 이때 임의의 단어군 $\{\boldsymbol{a}_k^i\}_{1\le k\le m, 1\le i\le T} \in \mathbf{R}^n$의 공통벡터는 식 11과 같이 구해진다.

$$\begin{aligned}\tilde{\boldsymbol{a}}_{common}^{i,Total} &= \tilde{\boldsymbol{a}}_{k,common}^{i} = \boldsymbol{a}_k^i - \bar{\boldsymbol{a}}_k^i \\ &= \boldsymbol{a}_1^i - \bar{\boldsymbol{a}}_1^i \ (k=1,2,\cdots,m)\end{aligned} \tag{11}$$

$$\bar{\boldsymbol{a}}_k^i = <\boldsymbol{a}_k^i, \boldsymbol{z}_1^i>\boldsymbol{z}_1^i + \cdots + <\boldsymbol{a}_k^i, \boldsymbol{z}_{m-1}^i>\boldsymbol{z}_{m-1}^i \tag{12}$$

단, $\boldsymbol{z}$는 $\boldsymbol{b}_i$ $(i=1,2,\cdots,m-1)$로부터 Gram-Schmidt의 직교화 방법을 사용하여 얻은 정규직교벡터로서 $\boldsymbol{b}_i$의 기저를 이룬다.

$$\begin{aligned}\boldsymbol{b}_1^i &= \boldsymbol{a}_2^i - \boldsymbol{a}_1^i \\ \boldsymbol{b}_2^i &= \boldsymbol{a}_3^i - \boldsymbol{a}_1^i \\ &\vdots \\ \boldsymbol{b}_{m-1}^i &= \boldsymbol{a}_m^i - \boldsymbol{a}_1^i\end{aligned} \tag{13}$$

이때 얻어진 $\tilde{\boldsymbol{a}}_{common}^{i,Total}$에 발산정보를 극대화시키기 위해 선형변환 행렬 $\boldsymbol{W}^T$를 곱해 주었다.

$$\boldsymbol{Y}_{common}^{i,Total} = \boldsymbol{W}^T \tilde{\boldsymbol{a}}_{common}^{i,Total} \tag{14}$$

위 식에서 $\boldsymbol{Y}_{common}^{i,Total}$는 m-차원 벡터, $\tilde{\boldsymbol{a}}_{common}^{i,Total}$는 n-차원 벡터, 그리고 $\boldsymbol{W}^T$는 각 행이 일차독립인 $m \times n$행렬이다. 각 단어군에서 얻어진 전체 공통벡터 $\tilde{\boldsymbol{a}}_{common}^{i,Total}$간의 within-class scatter값을 최대가 되도록, 전체 공통벡터들간의 공분산 행렬을 구한 후 고유치가 영에 근접하는 고유벡터들을 이용하여 선형 변환식 $\boldsymbol{W}$를 만들었다. 구해진 변환행렬 $\boldsymbol{W}^T$를 전체 공통벡터들에 곱해주면 식 15와 같이 되며, 전체 공통벡터간의 구분정보가 최대가 된다.

$$\begin{aligned}\boldsymbol{W}^T \tilde{\boldsymbol{a}}_{common}^{i,Total} &= \boldsymbol{W}^T \boldsymbol{a}_k^i - \boldsymbol{W}^T \bar{\boldsymbol{a}}_k^i \\ &= \boldsymbol{W}^T \boldsymbol{a}_k^i - \boldsymbol{W}^T \boldsymbol{Z}^i \boldsymbol{Z}^{iT} \boldsymbol{a}_k^i \ (k=1,2,\cdots,m)\end{aligned} \tag{15}$$

위 식에서 알 수 있듯이 $\boldsymbol{W}^T$는 전체 공통벡터 추출에 아무런 영향을 미치지 않으며, 단지 전체 공통벡터들간의 구분정보만 최대가 되도록 변환시킨다.

Ⅳ. 실 험

실험은 ETRI에서 제작한 음성 데이터베이스중 숫자음과 부서명을 사용하였으며, MATLAB을 이용하여 인식실험을 수행하였다[5].

사용된 단어들은 16 kHz의 표본화율을 가지며 16 bit

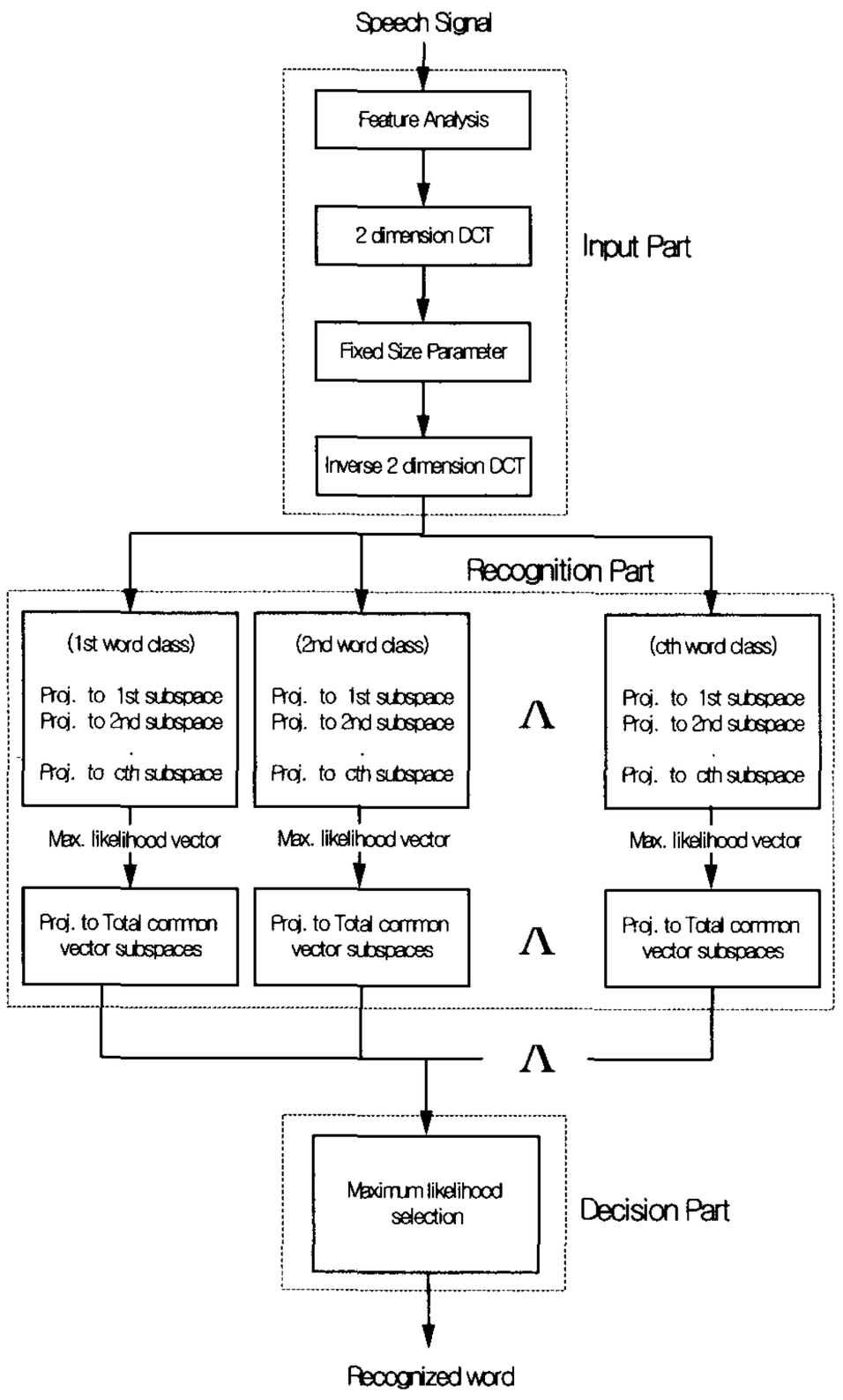

그림 2. 제안된 알고리즘을 이용한 음성인식 구성도
Fig. 2. Block diagram of proposed algorithm.

로 저장된 음성 데이터들이다. 음성신호는 특성추출을 위해 0.015초 간격으로 절단하였고 50%씩 중복하여 사용하였다. 그리고 Hamming 창함수를 사용하였으며 32차 청각필터를 사용하여 음성신호에서 특징 파라미터를 추출한 후, 얻어진 특징벡터들의 시간축 정규화를 위하여 2차원 DCT를 사용하였다[2][6]. 그리고 다양한 길이의 단어군들을 인식하기 위하여 벡터의 마지막 부분에 단어의 프레임 수를 추가하였다. 인식에 사용된 전체적인 구성도는 그림 2와 같다.

1. 훈련인원 변화에 따른 인식실험

실험에 사용된 숫자음 음성은 모두 3,200개이며 이중 각각 10명, 20명의 서로 다른 화자가 1회 발성한 숫자음을 학습 데이터로 사용하였다. 학습에 사용되지 않은 20명의 음성은 인식실험용으로 사용하였다. 그리고 부서명 음성은 모두 2,002개이며, 이중 각각 10명, 20명, 40명의 서로 다른 화자가 발성한 부서명 22개를 학습 데이터로 사용하였다. 모든 실험에서 학습용으로 선택된 화자들은 남성과 여성의 비율이 각각 50%씩 되도록 설정하였다. 실험은 모두 2가지 방법으로 진행되었다. 우선, 학습용 화자의 수를 10명과 20명으로 변화시켜가며 실험하였다. 두 번째로는 기존 방법을 이용한 경우와 KCSM만을 이용한 경우, 제안한 두 가지 방법을 모두 이용한 경우의 인식률 실험을 각각 수행하였다.

표 1. 고립 숫자음 인식률(%)
Table 1. Recognition rate of isolated digit.

• 벡터 크기 : 14(filter axis)×11(frame axis)

• k-clustering 개수 : 4

• 파라미터 : 32차 청각모델

인식 방법	CVEM	KCSM	KCSM+DCVEM
훈련 인원 10명	96.06	92.75	96.25
훈련 인원 20명	98.00	97.81	98.37

표 2. 고립단어 인식률(%)
Table 2. Recognition rate of isolated word.

• 벡터 크기 : 14(filter axis)×11(frame axis)

• k-clustering 개수 : 4

• 파라미터 : 32차 청각모델

인식 방법	CVEM	KCSM	KCSM+DCVEM
훈련 인원 10명	91.53	89.21	90.19
훈련 인원 20명	94.11	90.90	93.13
훈련 인원 40명	95.63	94.83	95.72

2. 단어군의 소그룹 개수 변화에 따른 인식실험

KCSM을 이용해 소그룹의 개수를 변화시키며 실험한 결과를 표 3, 4에 나타내었다. 고립 숫자음의 경우는 20명의 화자가 1회 발성한 숫자음 20개를 학습에 사용하였으며 인식실험에는 학습에 사용되지 않은 20명의 화자가 4회 발성한 숫자음 20개를 사용하였다. 부서명의 경우는 40명이 1회 발성한 22개 부서명을 학습에 사용하였고 인식실험에는 학습에 사용되지 않은 51명이 1회 발성한 22개의 부서명을 사용하였다.

표 1과 표 2의 실험결과를 살펴보면, 학습용 단어의 수가 증가할수록 KCSM과 CVEM은 유사한 인식률 보

표 3. 고립 숫자음 실험 결과
Table 3. The isolated digit experimental result.

군집화 개수	CVEM	2	4	6	8	9
인식률 (%)	98.12	98.50	98.50	98.25	97.25	97.50

표 4. 고립단어 실험결과
Table 4. The isolated word experimental result.

군집화 개수	CVEM	2	4	6	8	9
인식률 (%)	94.38	95.45	94.38	94.20	94.74	94.56

이며, KCSM과 DCVEM을 결합한 방법은 더 높은 결과를 보여주고 있다. 이러한 결과는 많은 학습용 단어가 사용될 경우 제안된 KCSM이 적용 가능하고 KCSM과 DCVEM을 결합한 방법이 우수함을 보여준다. 그리고 표 3와 표 4로부터는 분할되는 소그룹의 개수는 인식률에 큰 영향을 미치지 못하며, 기존 방법을 이용한 경우와 유사한 인식률을 얻음을 알 수 있다. 그러나 너무 많은 소그룹으로 학습용 벡터들을 분류할 경우 인식부분에서 많은 연산량을 필요로 하게 되므로 학습용 음성에서 얻어진 벡터의 차원과 학습시킬 음성의 총 개수를 참조하여 적절하게 군집화를 수행하는 것이 바람직하겠다.

V. 결 론

본 논문에서는 M. Bilginer 등이 제안한 공통벡터 추출방법의 문제점을 개선하기 위하여 k-clustering subspace method(KCSM)와 discriminant 공통벡터 추출방법(DCVEM)을 제안하였다. 그리고 제안한 방법들의 유효성을 입증하기 위하여 ETRI에서 배포한 한국어 고립 숫자음 3,200개와 ETRI의 부서명 2,002개를 사용해 인식실험을 수행하였다. 실험 결과 k-clustering subspace 방법은 계산량의 큰 증가 없이 기존 방법이 가진 학습 단어 수 제한이라는 문제점을 해결할 수 있었으며, 인식률 면에서도 기존 방법과 유사한 결과를 얻을 수 있었다. 그리고 discriminant 공통벡터 추출방법은 기존 방법에서 발생하기 쉬운 오인식의 원인을 사전에 제거할 수 있었으며 이로 인해 인식률에서 향상된 결과를 얻을 수 있었다. 두 가지 방법을 모두 결합한 경우는 기존에 비해 최고 0.37% 정도의 향상된 결과를 보여 주었다. 향후 제안된 알고리즘을 hybrid HMM과 결합하고 인식에는 다양한 파라미터들을 사용하여 좀 더 향상된 시스템을 구축해볼 계획이며, 또한 고립단어에서 연속음성으로의 확장도 시도해볼 계획이다.

참 고 문 헌

[1] M. Bilginer et al.,, "A novel approach to isolated word recognition", *Speech and Audio Processing., IEEE Trans.* Vol. 7, pp.620-628, 1999.

[2] 남명우, 박규홍, 정상국, 노승용, "선형판별분석과 공통벡터 추출방법을 이용한 음성인식", *한국음향학회지*, 제 20권, pp. 35-41, 2001.

[3] Davod C. Lay, *"Linear Algebra and its Applications"*, seconded., ADDISON-WESLEY, pp. 62-63, 2000.

[4] Saon, G., Padmanabhan, M., Gopinath, R., Chen, S., "Maximun likelihood discriminant feature spaces", *Acoustics, Speech, and Signal Processing, IEEE International Conference*, Vol. 2, pp. II1129 -II1132. 2000.

[5] Mathworks, *"Signal processing toolbox user's guide"*, Mathworks Inc. 1996.

[6] Lawrence Rabiner and Biing-Hwang Juang, *"Fundamentals of speech recognition"*, Prentice -Hall International, Inc., 1993.

저 자 소 개

남 명 우(정회원)

1992년 서울시립대학교
제어계측공학과 학사 졸업.

1994년 서울시립대학교
전자공학과 석사 졸업.

2001년 서울시립대학교 전자공학과
박사 졸업

2003년~현재 혜전대학 디지털전자디자인과 교수.

<주관심분야 : 음성인식, 신경망, 신호처리, 능동소음제어>